Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models

Tom Dreyfus, Valérie Doye, Frédéric Cazals

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models

Tom Dreyfus — Valérie Doye — Frédéric Cazals

N° 7768

October 2011

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--7768--FR+ENG

INRIA
SOPHIA ANTIPOLIS

Assessing the Reconstruction of Macro-molecular Assemblies with Toleranced Models

Tom Dreyfus [*] , Valérie Doye [†] , Frédéric Cazals [‡]

Thème BIO — Systèmes biologiques
Projet ABS

Rapport de recherche n° 7768 — October 2011 — 27 pages

Abstract: We introduce Toleranced Models (TOM), a generic and versatile framework meant to handle models of macro-molecular assemblies featuring uncertainties on the shapes and the positions of proteins. A TOM being a continuum of nested shapes, the inner (resp. outer) ones representing high (low) confidence regions, we present statistics to assess features of this continuum at multiple scales. While selected statistics target topological aspects (pairwise contacts, complexes involving proteins of prescribed types), others are of geometric nature (geometric accuracy of complexes).

We validate the TOM framework on recent average models of the Nuclear Pore Complex (NPC) obtained from reconstruction by data integration, and confront our statistics against experimental findings related to sub-complexes of the NPC.

In a broader perspective, the TOM framework should prove instrumental to handle uncertainties of various kind, in particular in electron-microscopy and crystallography.

Key-words: Macro-molecular assembles, Reconstruction by data integration, Nuclear Pore Complex, Model assessment, Fuzzy models, Toleranced Models, Curved Voronoi diagrams, curved α-shapes.

[*] INRIA Sophia-Antipolis-Méditerranée, Algorithms-Biology-Structure; Corresponding author: Tom.Dreyfus@sophia.inria.fr
[†] Institut Jacques Monod, CNRS, UMR 7592, Univ Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France
[‡] INRIA Sophia-Antipolis-Méditerranée, Algorithms-Biology-Structure; Corresponding author: Frederic.Cazals@sophia.inria.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Évaluation de la Reconstruction de Gros Assemblages Protéiques avec des Modèles Tolérancés

Résumé : Ce travail introduit le canevas des modèles tolérancés (TOM), afin de modéliser des assemblages macro-moléculaires présentant des incertitudes tant sur la forme des protéines que sur leur position. Un modèle tolérancé étant un continuum de formes emboîtées, nous présentons une panoplie de statistiques permettant d'évaluer ces formes à diverses échelles. Certaines des statistiques qualifient des aspects topologiques (contacts deux à deux, sous-complexes impliquant certaines protéines spécifiques), alors que d'autres sont de nature géométrique (taille des sous-complexes).

Nous validons le canevas en évaluant les modèles moyennés du pore nucléaire (NPC) reconstruits récemment par intégration de données, et confrontons nos statistiques à divers résultats relatifs à des sous-complexes.

De façon prospective, le canevas des modèles tolérances devrait permettre de gérer des incertitudes diverses, en particulier en cryo electron microscopie et crystallographie.

Mots-clés : Assemblages Macro-moléculaire, Reconstruction par intégration de données, Pore nucléaire, Evaluation de modèles, Modèles flous, Modèles tolérancés, Diagrammes de Voronoi courbes, α-shapes courbe.

Introduction

The reconstruction of large protein assemblies is a major challenge due to their plasticity but also the flexibility of the proteins involved. For assemblies involving of the order of tens of proteins, the reconstruction may be based on the docking of atomic resolution structures within cryo-electron microscopy maps [1]. But this strategy fails for systems whose complexity is one order of magnitude higher, especially if a number of atomic resolution crystal structures are missing, or if the resolution of the cryo-electron microscopy data is not sufficient. In that case, additional pieces of information must be resorted to, in particular biochemical data encoding spatial proximity within the assembly, such as Tandem Affinity Purification data. This reconstruction by data integration [2] aims at finding the model(s) best complying with the experiments, but uncertainties and inherent ambiguities (affinity purification data) on the input generally preclude a unique reconstruction. This in turn makes quantitative assessments with respect to the experimental data non trivial, let alone the mechanistic exploitation of the models.

The recent reconstruction of plausible models of the yeast Nuclear Pore Complex (NPC) illustrates this situation [3, 4]. The NPC, a protein assembly with eight-fold radial symmetry anchored in the nuclear envelope, regulates the nucleo-cytoplasmic transport. It is composed of $\sim$ 30 distinct proteins types each present in multiple copies, and is the largest protein assembly known to date in the eukaryotic cell [5, 6]. Using the aforementioned data integration approach, a set of 1000 plausible coarse-grain structures were optimized, and were averaged to compute probability density maps (maps for short in the sequel) of the individual protein species of the yeast NPC [4]. These maps present a prototypical example of uncertain data since placing the isolated protein instances within a map does not admit a unique solution. These ambiguities hindered the quantitative exploitation of the reconstructions, and motivate our developments.

In this work, we introduce Toleranced Models (TOM), a modeling framework derived from the theory of curved Voronoi diagrams [7], whose hallmark consists of replacing a fixed shape by a continuum of nested shapes. We first explain how TOM can be used to accommodates uncertainties, proceed with geometric and topological statistics characterizing the continuum, and by applying this machinery to the NPC, illustrate how TOM rapidly allow a quantitative assessment of models featuring uncertainties.

Results

Methodology

Toleranced Models, Hasse diagrams, and multi-scale analysis. Consider a protein assembly with uncertainties on the shape and/or the position of its constituting proteins. To deal with such uncertainties, the toleranced model framework consists of designing proteins as collections of contiguous toleranced balls. A toleranced ball $\overline{B_i}$ is a pair of concentric balls of radii r_i^- and r_i^+, the inner and outer balls, respectively meant to encode high confidence regions (radius r_i^-) and uncertain regions (radius r_i^+) (Fig. 1(A,B)). We next introduce the parameter $\lambda > 0$, governing a growth process consisting of linearly interpolating (or extrapolating if $\lambda > 1$) the radii. That is, the grown ball $\overline{B_i}[\lambda]$ stands for the ball of radius :

$$r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-). \tag{1}$$

For $\lambda = 0$ (resp. $\lambda = 1$), the grown ball matches the inner (resp. outer) ball.

A toleranced model being a continuum of shapes, we analyze it at multiple scales. For a given value of λ, we define a connected component of the domain covered by the grown proteins as a region such that there exists a path between any two points of this region. We expect these connected components to correspond to biochemical entities (proteins or complexes). Growing lambda triggers merges between the proteins, leading to complexes of increasing complexity. To track these merges, the Voronoi region of a toleranced ball $\overline{B_i}$ is defined as the set of points reached by $\overline{B_i}[\lambda]$ before any other ball $\overline{B_j}[\lambda]$, and the intersection between a grown ball and its Voronoi region is called a Voronoi restriction. A merge between two proteins exactly occur when two Voronoi restrictions, one from each protein, intersect on the bisector separating the Voronoi regions of their defining balls (Fig. 1(C)). The merges can be recorded in a direct acyclic graph called a Hasse diagram (black graph on Fig. 1(D)); its nodes are complexes, possibly reducing to a single protein, and its edges encode the merge of complexes along the growth process. Hasse diagrams can also be restricted to merge events involving protein instances whose types belong to a prescribed set T (red graph of Fig. 1(D)). In this bicolor setting, where the instances whose type is in T are called red and the remaining ones blue, a complex C is termed an isolated copy

if each type of the prescribed T set is present exactly once in C. For assemblies involving multiple instances of a given protein, the number and the lifetime of isolated copies provides a measure of the separability of the different copies of a complex involving all the types of the set T.

As explained in the Methods section, tracking merges through colliding restrictions, as opposed to tracking mere intersections between growing balls, is compulsory in the bicolor setting. But the fact that Voronoi bisectors are mathematically defined by degree four algebraic surfaces complicates matters, and a construction called the (partial) λ-complex, introduced in [8], must be resorted to (Methods and Supplemental).

The geometric accuracy of a complex along the growth process can be assessed by its volume ratio, defined as:

$$\overline{V}_\lambda(C) = Vol_\lambda(C)/Vol_{ref}(C), \tag{2}$$

where $Vol_\lambda(C)$ is defined as the sum of the volumes of the Voronoi restrictions of the balls constituting the complex, and $Vol_{ref}(C)$ is obtained by adding up the reference volume of each protein as estimated from its sequence [9].

Contact probabilities. At the local level, the complexes encountered in the Hasse diagram can be used to evaluate protein contacts with respect to 3D templates known at atomic resolution. To quantitatively characterize pairwise contacts between instances of two protein types (P_i, P_j), we define a contact probability depending on the stoechiometry k of the interaction between these two proteins by

$$p_{ij}^{(k)} = 1 - \lambda(P_i, P_j)/\lambda_{\max}, \tag{3}$$

with $\lambda(P_i, P_j)$ the first value of λ for which k contacts are established between instances of two protein types (P_i, P_j), and with $\lambda_{\max}$ the λ value where the growth process stops. As explained in the Methods section, $\lambda_{\max}$ is set so that the volume ratio of Eq. (2) matches the uncertainties observed in the input data (details in Methods).

The variation of $p_{ij}^{(k)}$ as a function of k, called the contact curve, is a key feature to assess whether an unambiguous stoichiometry exists for the contact between instances of two types. We use this contact curve to define: (i) k_{high}, the largest stoichiometry observed for the probability $p_{ij}^{(1)}$; (ii) k_{low} as the largest stoichiometry for which $p_{ij}^{(k)} > 0$; (iii) k_{drop}, the stoichiometry maximizing the probability drop $\delta p_{ij}^{(k)} = p_{ij}^{(k)} - p_{ij}^{(k+1)}$; (iv) $s(k_{drop}) = p_{ij}^{(1)}/\delta p_{ij}^{(k_{high})}$ the significance of the largest variation with respect to $p_{ij}^{(1)}$.

Validation

We validate the previous method on the NPC model of [3], which involves 30 types whence 34 maps due to four duplicated types (Nup82, Nsp1, Nic96, Nup145N). The map of Gle1 being missing from the *localization volumes* section of `http://salilab.org/npc/`, we use the remaining 33 maps as input, for a total of 29 types. The eight-fold axial symmetry of the NPC and the presence of 28 and 29 instances in the cytoplasmic and nuclear half-spokes account for a total of $8 \times (28 + 29) = 456$ instances.

Building a TOM for the NPC. Given a map, we assigned to each protein instance an occupancy volume, i.e. a set of voxels within the map, using the strategy described in [3]. We covered each such volume with 18 toleranced balls of identical radius, whose organization into distinct canonical shapes was dictated by the morphology of the volume (Fig. 1(E); details in Methods). The inner radii of the balls were set so that the volume of each instance matches the reference volume estimated from the protein sequence [9]. The outer radii were set such that for $\lambda = 1$ the volumes of all instances in the assembly match the uncertainties of the input data (details in Methods). Merging the TOM of all maps yields the TOM of the whole NPC (Fig. 1(F)).

Since these proteins were present in multiple copies in this assembly (stoichiometry of 8, 16 or 32), different shapes could be attributed to distinct instances, a feature that can be used to assess symmetries within an assembly (Fig. 1(F) and supplemental table 2). Indeed, analyzing the shapes taken by the various instances of a given type in the toleranced model of the NPC revealed some variations featuring a non-optimal symmetrical organization. In addition, while their overall shape was generally consistent with the topologies defined in [4], notable exception was noticed for a subset of nucleoporins characterized by FG (phenylalanine-glycine) repeated sequences (FG-Nups) (supplemental Table I) , a feature likely reflecting the highly flexible structure of these repeated sequences [10].

Contact frequencies f_{ij} between two types P_i and P_j were previously defined by [3] as the fraction of structures in the ensemble (of size 1000) for which at least one contact between these two proteins is observed (see the Contact Frequency Information at `http://salilab.org/npc/`). However, these values did not reflect the stoichiometry of these interactions within the entire assembly. To overcome this limit, pairwise contact probabilities established from our toleranced model were systematically generated and are available from `http://cgal.inria.fr/abs/voratom`.

Analysis of the Y-complex. As a paradigm, to investigate the coherence / differences between the toleranced model and contacts reported in the literature, we used the Y-complex, an extensively-characterized heptamer composed of Nup133, Nup84, Nup145C, Sec13, Nup120, Nup85 and Seh1. Three-dimensional electron-microscopy of in vitro reconstituted Y complexes, followed by docking of nucleoporin crystal structures into these electron-microscopy maps previously revealed the arrangement of its seven subunit into a Y-shaped complex, involving 6 well-defined binary contacts between its members (Fig. 2(Aa)). Moreover, while the organization of the Y-complexes within the NPC had long been debated [11, 12], Kampman et al [13] recently provided a strong support for an head to tail arrangement of these complexes, in which 8 Y-complexes lie with their long axes parallel to the nuclear envelope plane and form two rings through interactions of Nup133 with the arms of the neighboring Ys (Fig. 2(Ab)).

The released maps [3] revealed an expected stoichiometry of 16 connected components for all members of the Y-complex, with the exception of Sec13, reflecting a more ambiguous position of the latter within this model (Fig. 2(Ac) and supplemental Fig. 3).

At the assembly level, analysis of the Y-complex using the Hasse diagram derived from our toleranced model was globally consistent with its expected organization as growing λ leads to the formation of two distinct entities, each corresponding to an individual ring (Fig. 2(B)). However, only 11 isolated copies appeared on this complete diagram. In contrast, analysis of either the Nup133-Nup84-Nup145C (Y-edge+Y-tail minus Sec13) or Nup120-Nup85-Seh1 (Y-arms) revealed 16 connected components (supplemental Fig. 4), reflecting in a quantitative manner the dissociation of the Y-complex into two distinct entities as visualized in the maps (Fig. 2(Ad)).

At the local level, previously established contact frequencies f_{ij} between the various types present in the Y-complex were not discriminative [3] (Fig. 2(Dc)). In contrast, contact probability analyses revealed that out of the 6 expected binary contacts within the NPC, 4 had a high probability to occur 16 times as expected ($k_{drop} = 16$) and one was slightly less consistent ($k_{drop} = 12$). However only 2 contacts were observed between Nup120 and Nup145C whereas additional pairs had an unexpected high contact probability (Fig. 2(Dc)), indicating that these proteins are poorly positioned with respect to each other in the current model. Finally, the implication of Nup133 in the ring closure was validated since Hasse diagrams of the Y-complex without Nup133 prevented the formation of two distinct entities (Fig. 2(C)). However, while it was previously suggested that interaction between Nup133 and Nup120 was required for ring closure [12], contact analysis only revealed 1 significant contact between these two proteins with however 6 additional contacts between Nup133 and Nup85.

A similar analysis on the Nsp1-containing complexes (supplemental Fig. 5) likewise highlights the coherence and the differences between data gathered by various laboratories in the NPC field, and quantitative assessments based on the TOM model.

Discussion

The core idea underlying the TOM framework is to replace a unique possibly ambiguous shape by a continuum of nested shapes encoding high and low confidence regions, so as to infer a finite set of events characterizing this continuum at multiples scales. As evidenced by our analysis of sub-complexes of the NPC, these events encode topological and geometric properties of the model, and allow its assessment at the local and assembly levels, by qualifying the coherence and the discrepancies against data gathered from the literature. Methods to assess the stoichiometry of contacts and the geometric accuracy of sub-complexes were not previously available. They are inherently provided by the TOM framework, so that our toolbox may be used within a virtuous loop *model reconstruction - model selection*, to promote the reconstructions which best comply with experimental data, including those obtained in other model organisms [14]. Also, since these statistics can be immediately gathered for any subset of proteins, the toolbox can be used to study and simulate any complex.

A key feature of the TOM framework is its versatility and genericity.

The framework is versatile with respect to the design of toleranced models. While simple canonical shapes to represent protein instances have been used in this study, the advent of new structural data calls for more elaborate representation schemes. An appealing strategy would be to accommodate the domains and the linkers of a protein with toleranced balls of different properties. Also, multi-scale TOM may be used to switch between representations at the domain, secondary structure, and atomic levels. Such representations would clearly provide a finer encoding of high and low confidence regions. However, their design is challenging since geometric covering problems (where one wishes to use as few geometric primitives as possible to cover a prescribed domain) are typically intractable (NP-complete problems), although effective solutions may be found for parsimonious representations.

The framework is also generic with respect to the growth model used to define the continuum of shapes. In this work, the linear interpolation of the radii of a toleranced ball warrants a so-called compoundly weighted distance, in conjunction with the eponym Voronoi diagram. But the growth process associated to any Voronoi diagram can be accommodated, and anisotropic Voronoi diagrams should prove adequate to handle anisotropic uncertainties. However, the computation of such diagrams and their (curved) α-shape are open problems, both from the combinatorial standpoint (inferring the intrinsic complexity of these diagrams) and from the numerical standpoint (mastering the numerics characterizing the merge events along the growth process).

In addition to the relevance in the context of reconstruction by data integration, the TOM framework has potentially numerous applications. In cryo-electron microscopy, the extensions just alluded to should prove interesting to model density maps with low signal-to-noise ratio. In crystallography, TOM could be used to encode anisotropic temperature factors, thus yielding models encompassing the classical Van der Waals and solvent accessible models. In molecular modeling finally, TOM could be used to blur static atomic resolution models, so as to encode flexibility related properties in the context of flexible docking.

Beyond biophysics, the TOM framework should also prove useful in engineering (material sciences), computer science (geometric modeling) or applied mathematics, since situations where uncertainties must be dealt with are commonplace.

Methods

Toleranced models: Theory

The toleranced models are actually tightly coupled to the theory of curved Voronoi diagrams and α-shapes. The following remarks are meant to intuitively clarify these constructions, and the reader is referred to [8] for the precise mathematical statements.

Toleranced balls: Inner and Outer Radii versus Interpolation and Extrapolation. Intuitively speaking, toleranced models are best described in terms of inner and outer balls, the elementary geometric operation consisting of interpolating the radius between these radii. But the radius can also be extrapolated on both ends. To see how, consider a toleranced ball $\overline{B_i}(c_i; r_i^-, r_i^+)$, centered at c_i and with radii $r_i^+ > r_i^-$, and let x be any point in 3D space. As shown in [8], the following radius interpolation

$$r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-) = \| c_i - x \|, \tag{4}$$

yields the so-called *compoundly weighted distance*:

$$\lambda(B_i, x) = \frac{1}{r_i^+ - r_i^-}(\| c_i - x \| - r_i^-). \tag{5}$$

Phrased differently, interpolating or extrapolating the radius is equivalent to varying the generalized distance of Eq. (5).

Compoundly Weighted Voronoi Diagrams. The locii of 3D points having ball $\overline{B_i}$ as nearest neighbor according to the generalized distance of Eq. (5) defines the Voronoi region of that ball:

$$\text{Vor\_region}(\overline{B_i}) = \{x \mid \lambda(B_i, x) \leq \lambda(B_j, x), \forall j \neq i\}. \tag{6}$$

The collection of all such regions defines the so-called *compoundly weighted Voronoi diagram*, or CW VD for short [7]. The bisectors of this diagram are degree four algebraic surfaces, and the dual complex of this diagram, namely the equivalent of the celebrated Delaunay and regular triangulations [15], is an abstract simplicial complex [8]. Equating two equations (6) shows that the bisectors of such a diagram are degree four algebraic surfaces, which can be bounded or unbounded [8].

The λ-complex and the partial λ-complex. We are interested in tracking intersections between restrictions—recall that a restriction is the intersection between a grown ball and its Voronoi region. For a given λ, the intersecting restrictions form a subset of all pairs of intersecting balls, and for a large enough value of λ, the pairs obtained are the abstract one-dimensional simplices of the dual complex of the CW VD. In the classical affine case, intersections between restrictions yield the so-called α-complex [16], which is a subset of the regular triangulation [15]. The α-complex generalizes to the CW VD, this curved α-complex being called the λ-complex in [8]. In particular, detecting intersections between restrictions, a particular type of event in the λ-complex, requires evaluating degree four polynomials at degree four algebraic numbers i.e. numbers which are themselves roots of a degree four polynomial [8].

Recall that in the bicolor setting, all protein types are split into red and blue types. Resorting to the λ-complex to identify complexes of a given color, as opposed to tracking pairs of intersecting balls, is actually compulsory in the bicolor setting. To see why, term the intersection point x between the red spheres bounding two balls $\overline{B_i}[\lambda]$ and $\overline{B_j}[\lambda]$ of *pure* if x is not contained within a blue ball. Denoting $\lambda(B_i, x)$ the compoundly-weighted distance between a point x and the ball $\overline{B_i}$, pure intersections correspond to privileged binary contacts in the following sense: a pure contact point x is such that

$$\lambda(B_i, x) = \lambda(B_j, x) \leq \lambda(B_k, x), \forall k \neq i, j. \tag{7}$$

Contacts between restrictions retrieved from the λ-complex have this property. On the other hand, contacts directly read from pairs of intersecting balls may not be pure. For example, on Fig. 1, the first intersection point between p_1 and p_3 is contained within a blue ball. The λ-complex allows one to report all pure intersections, which are of two types, depending on whether the $\leq$ condition of Eq. (7) is taken in a strict sense or not: a strict condition corresponds to the so-called Gabriel edges of the λ-complex, while an equality corresponds to non-Gabriel edges (Supplemental Fig. (1). We call the collection of Gabriel edges (and more generally of Gabriel simplices) the *partial* λ-complex. The incentive for introducing the partial λ-complex is of computational nature: the only known algorithm to compute the λ-complex has $O(n^5)$ complexity, while a naive scan of all pairs of toleranced balls yields a computation of Gabriel edges in $O(n^3)$ time—with n the number of toleranced balls [8].

To assess these complexities, recall that our model contains 29 and 27 protein instances in the cytoplasmic and nuclear half-spokes, respectively. With 18 balls per toleranced protein, we get a total of $8 \times (29+27) \times 18 = 8064$ toleranced balls.

These sizes and complexities explain why the λ-complex can be computed on a half-spoke, but not on the whole NPC, which we processed with the partial λ-complex. However, as explained in the supplement, the difference between both complexes is not significant.

Constructing Toleranced Models

As sketched in the Results section, we build a toleranced model for the NPC based on the maps of the 30 protein types computed in [3]. More precisely, we build a toleranced model for each protein type from its map, and merge the toleranced models of all types to obtain the toleranced model of the whole NPC. In the sequel, we therefore focus on the processing of a given map.

From Occupancy Volumes to Canonical Protein Shapes. Processing a given map is a three-stage process. First, we allocate occupancy volumes to protein instances. This step consists of collecting voxels in such a way that the volume covered by these voxels matches the estimated volume of all instances, namely Vol_{ref} multiplied by the stoichiometry of the type. These voxels are collected by a greedy region growing strategy, as explained in [3, Caption of Fig.9]. Second, we compute a canonical shape involving 18 toleranced balls for each instance, out of four canonical shapes: linear, semi-linear, flat and roughly isotropic. Three of these shapes are illustrated on Fig. 1(E), the linear shape being omitted since we found that it does not represent any instance. (To compare, recall that in [3], at most 13 balls are used to represent a protein instance.) To see how a canonical shape is assigned, consider an occupancy volume O_V to be covered with 18 toleranced balls of identical radius. We perform a principal component analysis (PCA) of the centers of the voxels in O_V, from which we derive three couples (eigen value, eigen vector), denoted $(v_1(O_V), e_1(O_V))$, $(v_2(O_V), e_2(O_V))$ and $(v_3(O_V), e_3(O_V))$. Consider now the three couples obtained from the PCA of the centers of a canonical configuration, denoted $(v_1(C_S), e_1(C_S))$, $(v_2(C_S), e_2(C_S))$ and $(v_3(C_S), e_3(C_S))$. Let σ be a permutation of the symmetric group of size 3—there are 6 such permutations. We determine which canonical shape best represents the volume O_V by

picking the permutation minimizing the following sum:

$$\Sigma_{i=1}^{i=3}(v_{\sigma(i)}(C_S)e_{\sigma(i)}(C_S) - v_i(O_V)e_i(O_V))^2 \tag{8}$$

For each protein type, the number of instances for the four canonical shapes in the TOM of the NPC is provided in supplemental table 2.

Setting the Inner and Outer Radii. For a given canonical shape, the inner radius is set so that the volume of the union of the 18 inner balls matches the estimated volume of the protein Vol_{ref}. Since the maps of selected small proteins are less accurate than those of large proteins (supplemental Fig. 3), we set the outer radius such that the discrepancy $r_i^+ - r_i^-$ is proportional to α/r_i^-:

$$r_i^+ = \frac{\alpha}{r_i^-} + r_i^-. \tag{9}$$

Consider a collection of toleranced balls whose outer radii are set this way, that is $\{\overline{B_i}(c_i; r_i^-; r_i^+ = \frac{\alpha}{r_i^-} + r_i^-)\}$. Under the assumption $r_i^+ = \alpha/r_i^- + r_i^-$, the equation (5) becomes

$$\lambda(B_i, p) = \frac{r_i^-}{\alpha}(|| \, c_i - x \, || - r_i^-). \tag{10}$$

If one equates two such equations to define a Voronoi bisector, that is $\lambda(B_i, x) = \lambda(B_j, x)$, the α cancel out. Phrased differently, the CW VD of the toleranced balls does not depend on α. Therefore, we arbitrarily set $\alpha = 10$ and compute the whole λ-complex of the toleranced model.

To retain models with decent geometric accuracy, we set $\lambda_{\max} = 1$, a value such that the smallest volume ratio amongst all complexes, as defined by Eq. (2), is larger than 7:

$$\min_{\text{all complexes } C \text{ existing at } \lambda_{\max}} \overline{V}_{\lambda_{\max}}(C) \geq 7. \tag{11}$$

As seen from the supplemental Fig. 3, this value of 7 is the worst uncertainty observed over the input maps. This way, the end of the growth of the toleranced model is coupled to the uncertainties of the maps, measured in terms of volume ratios. We also note that for all the sub-complexes studied, all merge events occur before $\lambda_{\max}$. In particular, for the Y-complex, the last event triggering a topological change occurs at $\lambda = 0.68$, all the subsequent events corresponding to new contacts within the two rings.

About Volume Ratios. The volume ratio of Eq. (2) is meant to compare the volume of an instance to its reference volume estimated from its sequence, and we have just seen that the inner radii are set according to this reference volume.

However, it should be noticed that the inner radius value comes from a stand-alone computation—the canonical shape is not in contact with any other protein instance, while the volume ratio of Eq. (2) is evaluated for a protein instance within the toleranced model of the NPC. Therefore, because of overlaps between protein instances of different types, one may observe volume ratios < 1. This happens in particular when a given instance is surrounded by other instances featuring larger uncertainties, as the Voronoi regions of the balls of such instances tend to be large—whence reducing the Voronoi regions and the restrictions of the protein of interest.

Artwork

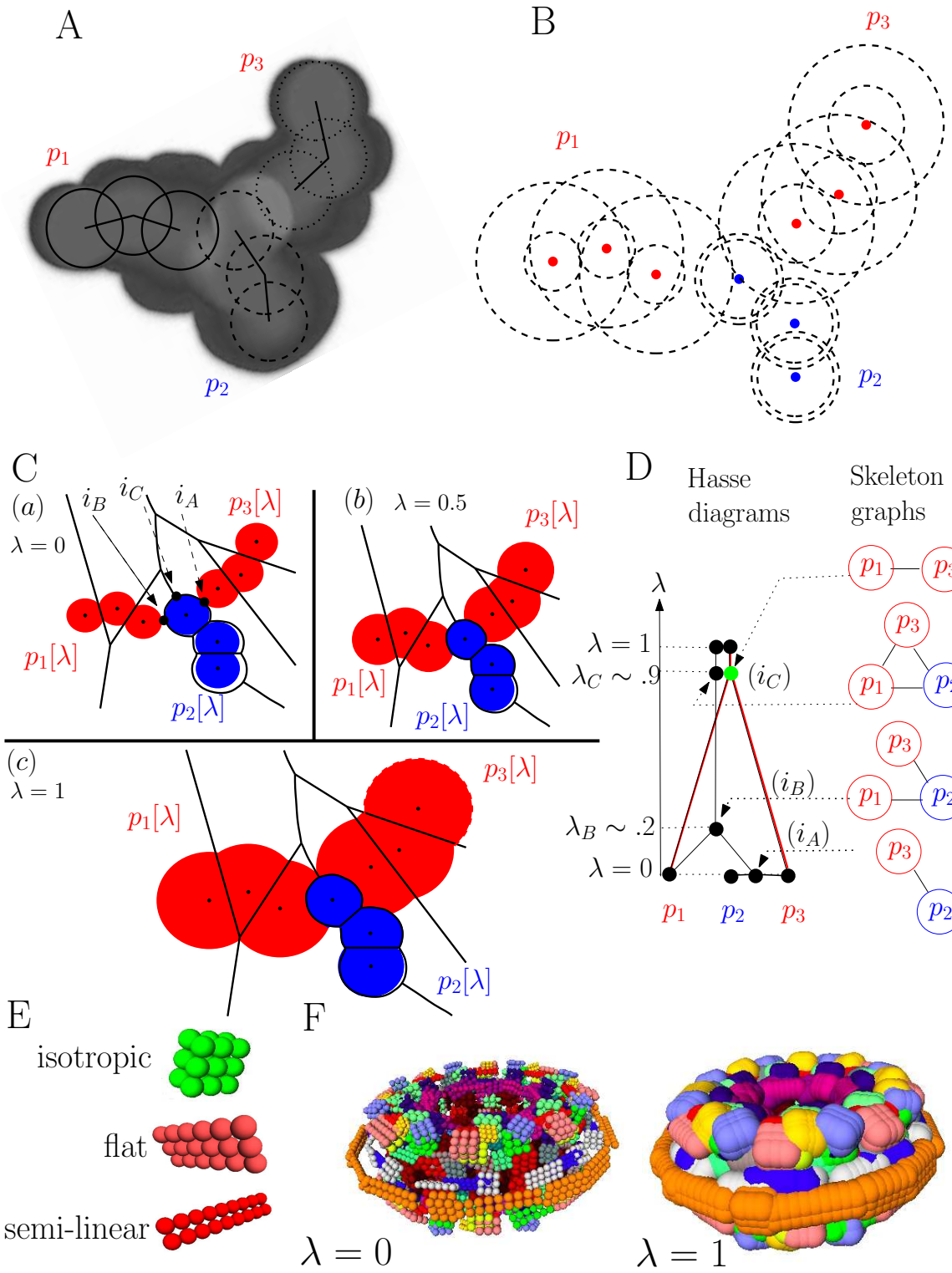
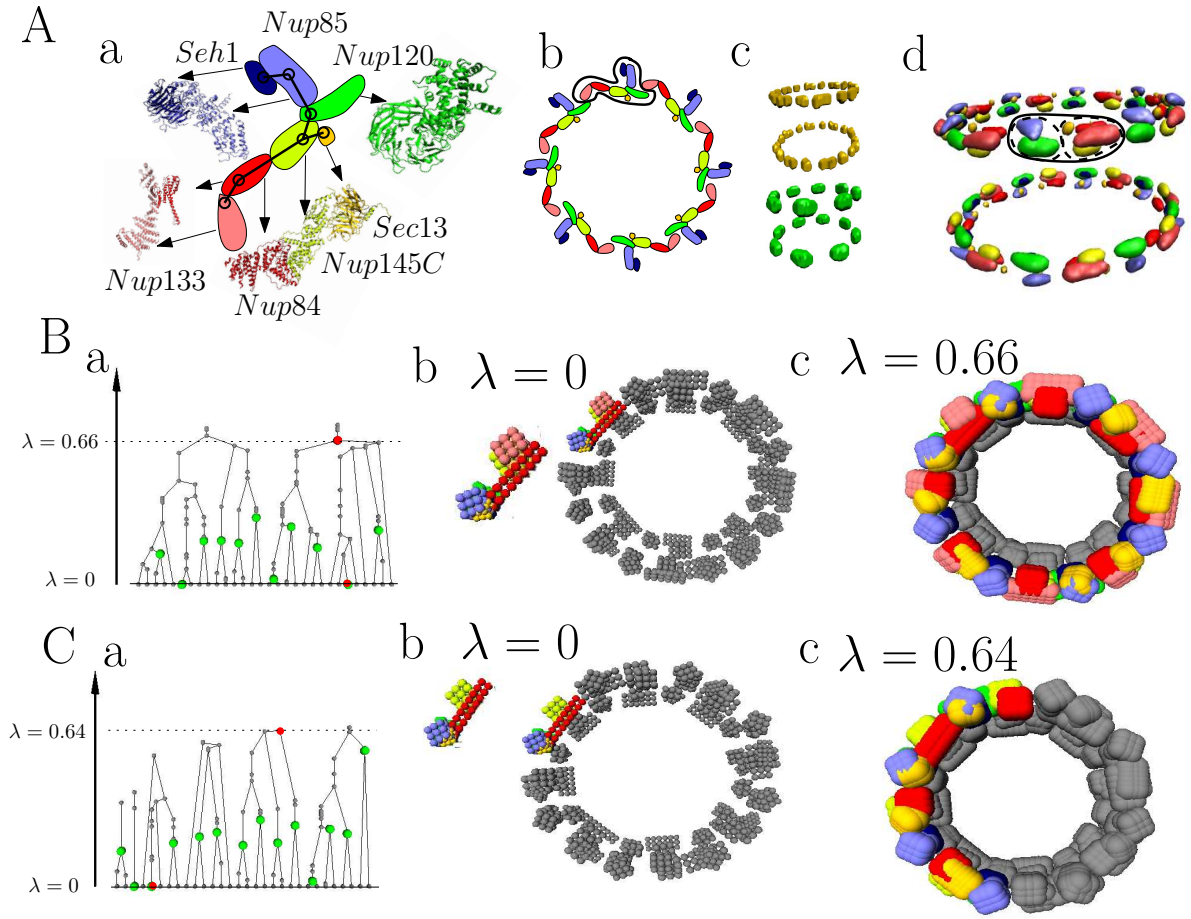


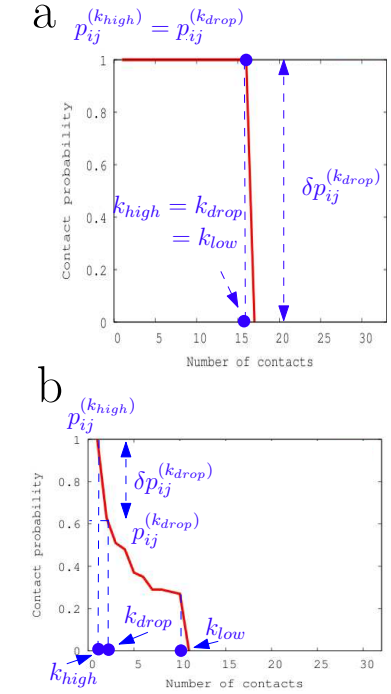
Figure 1:

Caption of Fig. 1

Tracking the interactions of three toleranced proteins. **A.** Conformations of three flexible molecules, and a map whose color indicates the probability of a given point to be covered by a random conformation of the ternary complex—from low (dark gray pixels) to high (light gray pixels) probabilities. **B.** The associated toleranced model, with three toleranced proteins, each defined by three toleranced balls. **C.** The bicolor toleranced model of Fig. (B), where it is assumed that each molecule corresponds to a distinct protein type; types of p_1 and p_3 belong to a prescribed set T, defining red types. Sub-figures (a,b,c) respectively show grown balls $\overline{B_i}[\lambda]$ for $\lambda = 0, 0.5, 1$. The region of the plane consisting of points first reached by a growing ball is the Voronoi region of this ball, represented by solid lines. Colored solid regions feature the *Voronoi restrictions* i.e. the intersection of a growing ball and its Voronoi region. Along the growth process, intersections between restrictions of two colors occur at the points i_A, i_B, i_C. **D.** Hasse diagrams encoding contacts between the proteins. Black graph: all proteins; red graph: red proteins only. On the latter, the green node corresponds to an isolated copy, i.e. a protein complex involving exactly one protein of the prescribed set T. **E.** The three different canonical shapes used in the toleranced model, of 18 balls each. **F.** The toleranced model of the NPC at $\lambda = 0$(**left**) and $\lambda = 1$ (**right**).



D



c

Protein types	f_{ij}	k_{high}	k_{drop}	$p_{ij}^{(k_{drop})}$	$s(k_{drop})$	$\min \overline{V}_{\lambda_{k_{high}}}$	$\max \overline{V}_{\lambda_{k_{drop}}}$
(Nup133, Nup84)	0.571	16	16	1.00	1.00	0.77	0.91
(Nup145C, Nup84)	1.000	16	16	1.00	1.00	0.77	0.88
(Nup120, Seh1)	0.837	16	16	1.00	1.00	0.82	0.92
(Nup133, Nup145C)	0.589	16	16	1.00	1.00	0.82	0.90
(Nup120, Nup85)	0.569	16	16	1.00	1.00	0.88	0.98
(Nup85, Seh1)	1.000	12	16	0.93	1.07	0.77	1.27
(Nup120, Sec13)	0.284	5	16	0.64	1.56	0.82	4.10
(Nup133, Sec13)	0.381	11	14	0.69	1.45	0.80	3.91
(Nup84, Sec13)	0.66	8	14	0.54	1.85	0.70	4.49
(Nup85, Sec13)	0.227	4	13	0.57	1.76	0.77	8.57
(Nup145C, Sec13)	0.503	12	12	1.00	1.00	0.79	0.86
(Sec13, Seh1)	0.233	4	9	0.65	1.55	0.57	6.88
(Nup120, Nup84)	0.487	1	8	0.60	1.68	0.91	3.26
(Nup84, Seh1)	0.376	1	7	0.49	2.06	0.79	3.89
(Nup133, Nup85)	0.478	1	6	0.79	2.34	1.28	2.65
(Nup145C, Seh1)	0.359	1	4	0.34	2.98	2.19	3.04
(Nup120, Nup145C)	0.498	1	2	0.86	2.21	1.02	1.49
(Nup120, Nup133)	0.465	1	1	1.00	2.18	0.92	0.92
(Nup84, Nup85)	0.543	1	1	1.00	2.71	0.89	0.89

Figure 2:

Caption of Fig. 2

Analysis of the the Y-complex. **A.** (a) Model of the Y-complex (adapted from [17]), with pairwise contacts (solid black lines) and the known crystal structures. (b) An embedding of the Y-complexes in the NPC (adapted from [17]). Eight instances of the Y-complex interact in a head-to-tail manner to form a ring. An instance of the Y-complex is circled in solid lines. (c) Maps of Sec13 (top) and Nup120 (bottom) (extracted from [4]); all voxels with a non null probability are shown for Sec13 (top) and Nup120 (bottom): while the first map is over-segmented, the second one contains the expected number of instances. (d) Level set surfaces of the seven maps—each level set is set to half of the maximum probability of the corresponding map. The solid circle singles out an instance of the Y-complex, while the dashed curves delimit two sub-units (Y-arm on the left; Y_X-edge + Nup133 on the right). **B.** (a) Hasse diagrams of the Y-complex with its seven types painted in red. The green nodes correspond to isolated copies, and the two red ones correspond to the colored complexes of (b) and (c). (b) Snapshot at $\lambda = 0$, with an isolated copy shown as inset. (c) Snapshot at $\lambda = 0.66$, when the upper ring appears. **C.** (a) Hasse diagrams of the Y-complex with the seven types but Nup133 painted in blue. The diagram now has six roots instead of two, which evidences the role of Nup133 in the closure of the two rings. (b) Snapshot at $\lambda = 0$, with an isolated copy shown as inset. (c) Snapshot at $\lambda = 0.64$, corresponding to one root of the Hasse diagram. The corresponding complex is a subset of the upper ring.
D. Contact curves and contact probabilities (a) Contact curve of (Nup84, Nup145C): 16 contacts are observed at $\lambda = 0$, and the contact probability is null for $k = 17$, which is the ideal situation since both types have a stoichiometry of 16. With a value of one, the significance coefficient $s(k_{drop})$ is also perfect, and the statistics on the volume ratios are excellent (Table Dc). (b) Contact curve of (Nup85, Nup84): The value $s(k_{drop}) = 2.71$ shows that the largest probability drop is not significant with respect to $p_{ij}^{(1)}$, and the large volume ratios evidence a poor positioning of the proteins— these ratios are larger than 3.56 for $k > k_{drop}$ which is equal to 1 in this case. In short, it is not possible to unambiguously choose a stoichiometry k for these two types. (c) Statistics summarizing contact curves: out of 21 pairs of the 7 protein types, 19 pair yield at least one binary complex. Pairs are sorted by decreasing k_{drop}, and are color-coded as follows: green: contacts of the skeleton of the Y-complex (Aa); red: putative contact accounting for the closure of the two rings [12]; orange: predominant contact accounting for the closure of the two rings in the TOM. The grey column displays the contact frequencies f_{ij} of [3].

Acknowledgements. Benoît Palancade (IJM, Paris, france) is acknowledged for helpful discussions.

References: Main Text

[1] W. Wriggers, R. A. Milligan, and J. A. McCammon. Situs: a Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy. *Journal of Structural Biology*, 125(2-3):185–195, 1999.

[2] F. Alber, F. Förster, D. Korkin, M. Topf, and A. Sali. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Ann. Rev. Biochem.*, 77:11.1–11.35, 2008.

[3] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, M.P. Rout, and A. Sali. Determining the Architectures of Macromolecular Assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[4] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, A. Sali, and M.P. Rout. The Molecular Architecture of the Nuclear Pore Complex. *Nature*, 450(7170):695–701, Nov 2007.

[5] S.R. Wente and M.P. Rout. The Nuclear Pore Complex and Nuclear Transport. *Colde Spring Harbor Perspectives in Biology*, 2(10):a000562, 2010.

[6] M.A. D'Angelo and M.W. Hetzer. Structure, Dynamics and Function of Nuclear Pore Complexes. *Trends Cell Biology*, 18:456–522, 2008.

[7] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams (2nd Ed.)*. Wiley, 2000.

[8] F. Cazals and T. Dreyfus. Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted α-shapes. In B. Levy and O. Sorkine, editors, *Symposium on Geometry Processing*, Lyon, 2010. Also as INRIA Tech report 7306.

[9] Y. Harpaz, M. Gerstein, and C. Chothia. Volume Changes on Protein Folding. *Structure*, 2:641–649, 1994.

[10] D.P. Denning, S.S. Patel, V. Uversky, A.L. Fink, and M. Rexach. Disorder in the Nuclear Pore Complex: the FG Repeat Regions of Nucleoporins are Natively Unfolded. *PNAS*, 100(5):2450–2455, 2003.

[11] S.G. Brohawn and T.U. Schwartz. Molecular Architecture of the Nup84–Nup145C–Sec13 Edge Element in the Nuclear Pore Complex Lattice. *Nat. Struct. Mol. Biol.*, 16(11):1173–1178, 2009.

[12] H.-S. Seo, Y. Ma, E.W. Debler, D. Wacker, S. Kutik, G. Blobel, and A. Hoelz. Structural and Functional Analysis of Nup120 Suggests Ring Formation of the Nup84 Complex. *PNAS*, 106(34):14281–14286, 2009.

[13] M. Kampmann and C.E. Atkinson and A.L. Mattheyses and S.M. Simon. Mapping the Orientation of Nuclear Pore Proteins in Living Cells with Polarized Fluorescence Microscopy. *Nat. Struct. Mol. Biol.*, 18(6):643–652, 2011.

[14] S. Amlacher, P. Sarges, D. Flemming, V. van Noort, R. Kunze, D. P. Devos, M. Arumugam, P. Bork, and E. Hurt. Insight into Structure and Assembly of the Nuclear Pore Complex by Utilizing the Genome of a Eukaryotic Thermophile. *Cell*, 146(2):277–289, 2011.

[15] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, UK, 1998. Translated by H. Brönnimann.

[16] H. Edelsbrunner. Weighted Alpha Shapes. Technical Report UIUCDCS-R-92-1760, Dept. Comput. Sci., Univ. Illinois, Urbana, IL, 1992.

[17] M. Kampmann and G. Blobel. Three-Dimensional Structure and Flexibility of a Membrane-Coating Module of the Nuclear Pore Complex. *Nat. Struct. Mol. Biol.*, 16(7):782–788, 2009.

1 Supplemental

1.1 Partial Computation of the λ-Complex

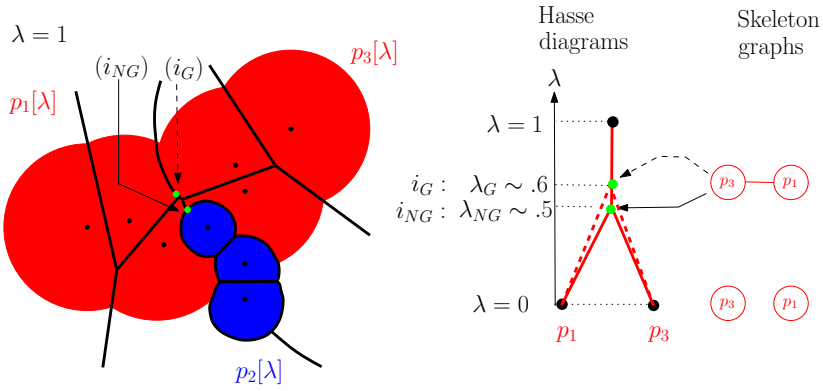
Partial versus whole λ-complex. As discussed in Methods, depending on the number of its constituting toleranced balls, a system may be investigated using the λ-complex or the partial λ-complex. Both algorithms were developed in C++, as discussed in Software, and the software was run on a dual core Intel Extreme CPU X7900 2.80GHz with RAM size of 8Go, under Fedora Core 14. The following running times were observed:

- The computation of the λ-complex on a complete half-spoke took about 15 hours. This computation for the whole NPC model halted after 6 days, due to a memory allocation failure.
- The computation of the partial λ-complex of the half-spoke and full NPC model respectively took less than one second and about 1 minute.

To assess the incidence of using the partial λ-complex rather then the λ-complex, recall that a contact between two proteins p_1 and p_2 corresponds to an edge of the λ-complex involving one ball of p_1 and one ball of p_2, and that such an edge has a status (Gabriel or not Gabriel). Using the partial complex may yield one of the following two discrepancies between p_1 and p_2 (supplemental Fig. 1):

- No edge connecting balls of p_1 and p_2 is Gabriel. In that case, the connexion between p_1 and p_2 is absent from the Hasse diagram derived from the partial λ-complex.
- There is at least one edge which is Gabriel, but this edge is encountered at $\lambda_G > \lambda_{NG}$, with λ_{NG} the value of λ corresponding to the first non Gabriel edge between balls of the two proteins. In that case, p_1 and p_2 are connected in the Hasse diagram derived from the partial λ-complex, but at λ_G.

In the following, we report statistics on these cases while working with one half-spoke of the NPC(594 toleranced balls for 33 protein instances).



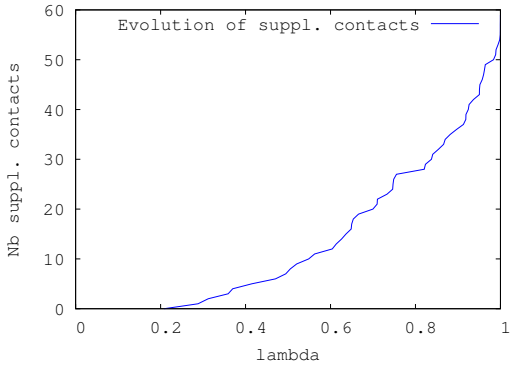
Supplemental Figure 1: λ-complex versus partial λ-complex. **Left.** A toleranced model of three proteins p_1, p_2, p_3 instantiated at $\lambda = 1$. Its associated compoundly weighted Voronoi diagram is drawn in solid black lines. The green points correspond to the first non Gabriel (i_{NG}) and first Gabriel (i_G) contact between the red proteins (p_1, p_3). **Right.** The Hasse diagram from the λ-complex (red solid lines) and the partial λ-complex (red dashed lines) restricted to the red proteins (p_1, p_3). For the λ-complex, the red proteins are connected at $\lambda_{NG} \sim 0.5$ at (i_{NG}). For the partial λ-complex, the red proteins are connected at $\lambda_G \sim 0.6$ at (i_G).

Missed protein contacts: global analysis. The supplemental Table 1 compares the number of edges and contacts for a half-spoke, from which it is seen that using the partial λ-complex yields a decrease of the number of edges and contacts of 62% and 31% percents, respectively. Moreover, as shown on the supplemental Fig. 2, the number of missed contacts increases slowly when λ increases, with only eight missed contacts for $\lambda < 0.5$.

Missed protein contacts: sub-complexes analyzed in this study. Regarding the protein contacts involved in the Y-complex, three differences between the two computations are observed at the level of one half-spoke :

	# edges	# contacts
Whole λ-complex	5947	193
Partial λ-complex	2227	133

Supplemental Table 1: λ-complex versus partial λ-complex on a half-spoke: comparison of the number of edges connecting toleranced balls, and of the number of contacts between protein instances.



Supplemental Figure 2: Evolution of the number of missed contacts between protein instances when using the partial λ-complex on a half-spoke, as a function of λ. See also the Supplemental Table 1.

– one contact (Nup85, Sec13) appears earlier, namely at $\lambda_{NG} = 0.39$ instead of $\lambda_G = 0.44$.
– two contacts are missed in the partial λ-complex: (Nup84, Seh1) at $\lambda = 0.88$ and (Nup120, Nup84) at $\lambda = 0.90$. Since the closure of the two rings is done at $\lambda = 0.64$, these two events are not relevant.

Concerning the T-complex and the Nup82-complex, there is no difference between protein contacts in both computations.

To conclude, using the partial λ-complex has no incidence on the results presented in this study, and makes the calculations tractable.

1.2 Probability Density Maps and Toleranced Models: Assessment

1.2.1 On the Probability Density Maps Used

The quality of the toleranced model depends on the accuracy of the maps used. In the following, on a per map basis, we report statistics aiming at qualifying these maps, in particular regarding the number of connected components (c.c.) of voxels having a non null probability, together with the volume of these c.c. with respect to the reference volume of the corresponding protein. (Following the terminology introduced for the volume ratio of Eq. (2), the reference volume $Vol_{ref}(P)$ of a protein P is the volume estimated from its sequence [18].)

On the Number of Connected Components. To discuss this statistic, the reader is referred to `http://cgal.inria.fr/abs/voratom` which contains snapshots of all the maps available from `http://salilab.org/npc/`.

Ideally, the number of c.c. of a map should match the stoichiometry of the corresponding protein. From the upper panel of the supplemental Fig. 3, one sees that this number is larger than / equal to / smaller than the stoichiometry in 5 / 19 / 9 cases. The former case corresponds to ambiguous locations which induce multiple connected components per instance, such as Sec13 with 32 c.c. for only 16 copies. The latter is due to the merge of nearby c.c. : the merging c.c. may be on the same side of the NPC but in two different spokes, such as Nsp1-1, or on the same spoke but on both sides of the NPC, such as Nup170. This phenomenon is extreme for Pom152, since a single c.c. corresponding to a filled torus is observed.

On the Volume of Connected Components. Assume that the map of the protein type P contains say n c.c.. Denoting $Vol(cc_i)$ the volume of the ith c.c., consider the set of map volume ratios

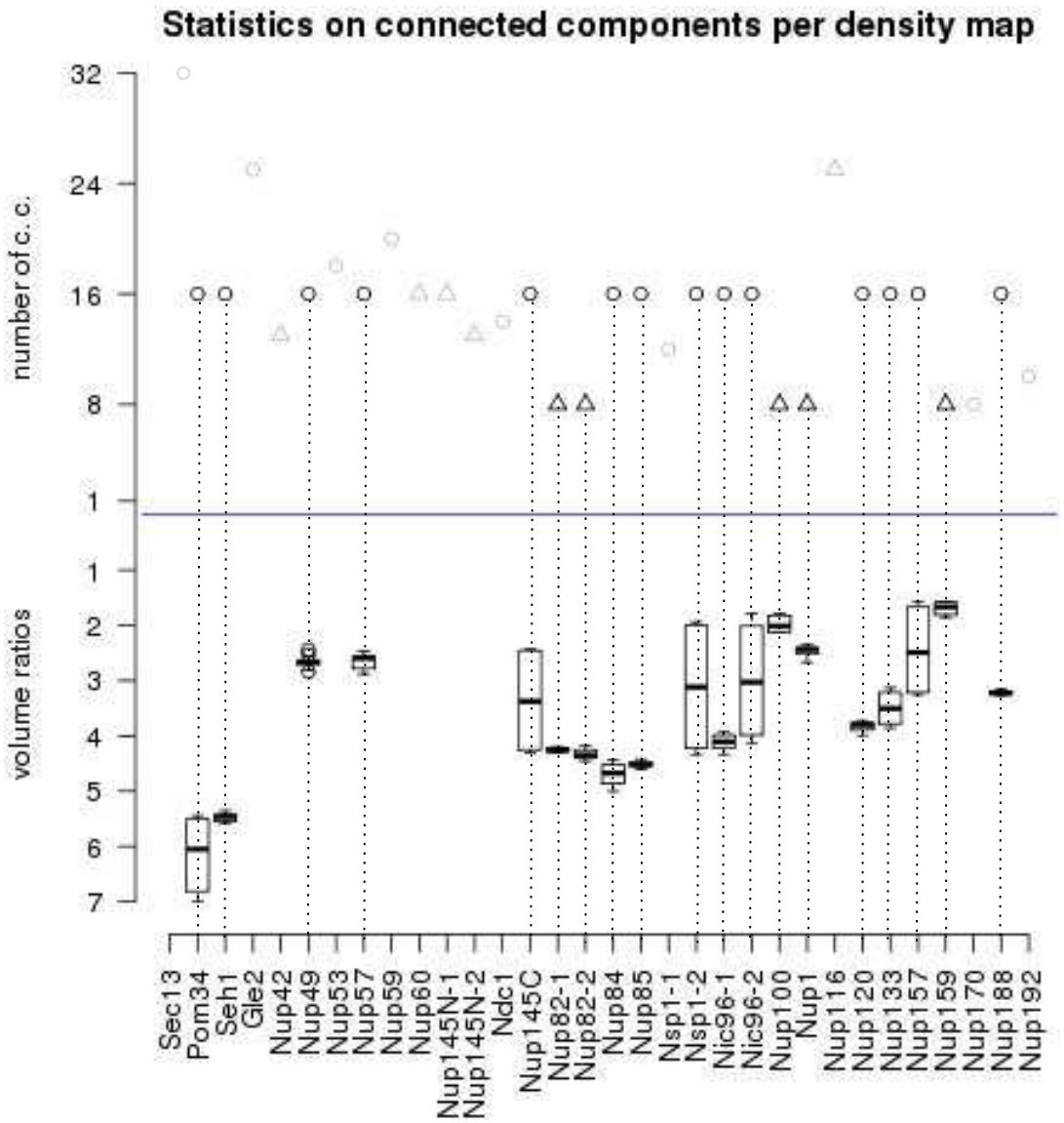
$$\overline{V}(cc_i) = Vol(cc_i)/Vol_{ref}(P), \text{ for } i = 1, \ldots, n. \tag{12}$$

Note that the map volume ratio is meant to assess the maps, as opposed to that of Eq. (2), which is geared towards the assessment of toleranced proteins.

The box plots [1] of the 19 maps with correct stoichiometry [2] are drawn on the supplemental Fig. 3(Lower-part).

[1] Recall that the box plot of a set of values is presented as follows. First, the rectangle displays three values, namely the first and third quartiles (small sides of the rectangle), and the median (bold line-segment inside the rectangle). Second, the whiskers extend to the extrema values of the plot, limited by 1.5 times the inter-quartile distance. Values below and above these thresholds are represented by circles.

[2] In this analysis, we restrict ourselves to maps which have the correct stoichiometry, since the meaning of c.c. in the remaining cases in unclear. For example, a c.c. within a plethoric map can be significant or can be insignificant. In theory, analyzing the relative importance of c.c. in any map can be done using Morse theory and persistence theory, in a manner similar to the algorithms developed in [19] in the context of Morse theory of the distance function. Yet, for general (density) maps, effective algorithms for Morse-Smale decompositions yet have to be developed.



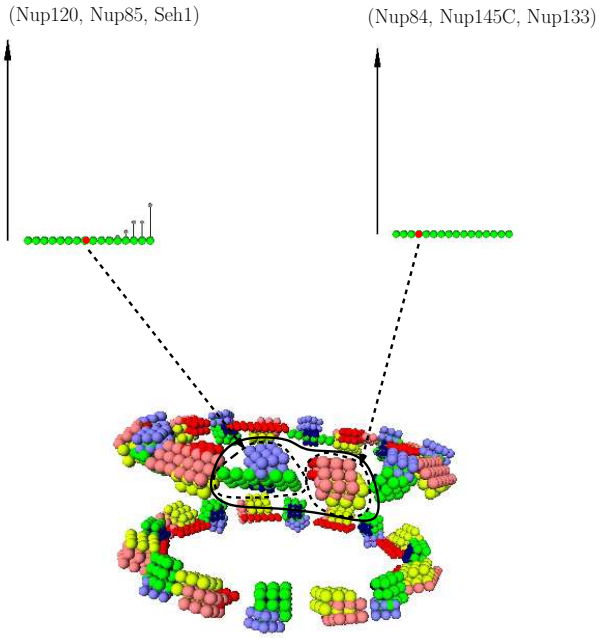
Supplemental Figure 3: Assessing the quality of the 32 maps—all maps but that of Pom152 which has a single connected component. The names of the 32 maps, with duplicate maps for Nup82, Nsp1, Nic96 and Nup145N, are displayed along the x-axis, and are sorted by increasing molecular weight (from 33.0×10^3 for Sec13 to 191.5×10^3 for Nup192, see the supplemental table 2). **Upper-part.** Number of connected components of voxels with non null density per map. Disks correspond to maps with a stoichiometry of 16, while triangles correspond to a stoichiometry of 8. The 19 maps with black marks exhibit the expected stoichiometry, as opposed to the 13 maps with grey marks. **Lower-part.** Box plots of the map volume ratios $\overline{V}(cc_i)$ of Eq. (12), for maps with a number of c.c. matching the expected stoichiometry of the protein type.

Protein type	Average Mol. Weight ($\times 10^3$)	Stoich.	#linear	#semi linear	#flat	#roughly isotropic	#balls in [4]
Nup192	191.5	16	0	0	0	16	2
Nup188	188.6	16	0	0	0	16	2
Nup170	169.5	16	0	0	7	9	2
Nup159*	158.9	8	0	0	4	4	11
Nup157	156.6	16	0	0	0	16	3
Pom152	151.7	16	0	6	8	2	10
Nup133	133.3	16	0	0	6	10	2
Nup120	120.4	16	0	0	2	14	2
Nup116*	116.2	8	0	0	0	8	13
Nup1*	113.6	8	0	0	4	4	9
Nup100*	100.0	8	0	0	0	8	13
Nic96-1	96.2	16	0	0	0	16	2
Nic96-2	96.2	16	0	0	0	16	2
Nsp1-1*	86.5	16	0	0	10	6	12
Nsp1-2*	86.5	16	0	0	0	16	12
Nup85	84.9	16	0	0	0	16	3
Nup84	83.6	16	0	1	5	10	3
Nup82-1	82.1	8	0	0	1	7	2
Nup82-2	82.1	8	0	0	0	8	2
Nup145C	81.1	16	0	0	0	16	2
Ndc1	74.1	16	0	1	7	8	2
Nup145N-1*	64.6	8	0	0	0	8	6
Nup145N-2*	64.6	8	0	0	0	8	6
Nup60*	59.0	8	0	0	0	8	4
Nup59*	58.8	16	0	0	6	10	4
Nup57*	57.5	16	0	0	0	16	3
Nup53*	52.6	16	0	0	6	10	3
Nup49*	49.1	16	0	0	0	16	3
Nup42*	42.8	8	0	0	3	5	5
Gle2	40.5	16	0	0	1	15	1
Seh1	39.1	16	0	0	0	16	1
Pom34	34.2	16	0	1	13	2	3
Sec13	33.0	16	0	0	4	12	1
Total	NA	448	0	9	86	353	NA

Supplemental Table 2: Protein types sorted by decreasing average molecular weights (dimensionless, first column), their expected stoichiometry (2nd column), the number of instances for the four canonical shapes in the toleranced model of the NPC (columns 3-6), and the number of balls used at the finest representation level by Alber et al. FG-Nups are denoted by $*$.

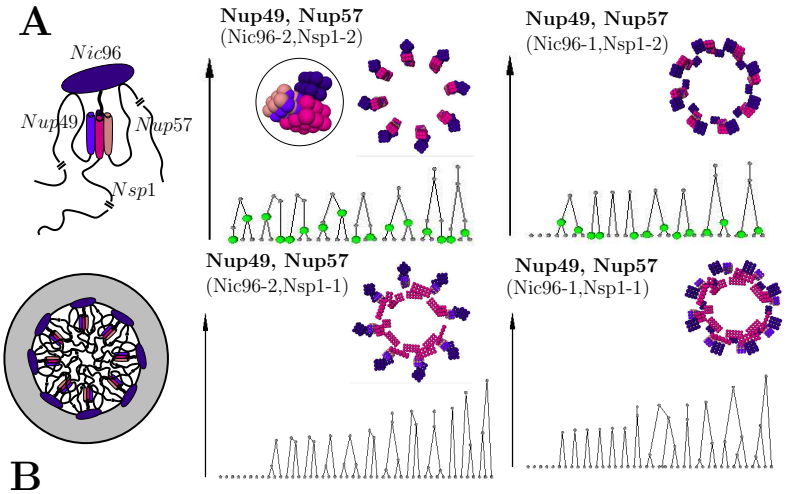
1.3 Further In-silico experiments on Sub-complexes of the Y-complex

On Fig. 1(Ad), we qualitatively observed a split of the Y-complex into two sub-complexes. The observation is substantiated by the 16 isolated copies observed on the two Hasse diagrams of the supplemental Fig. 4.

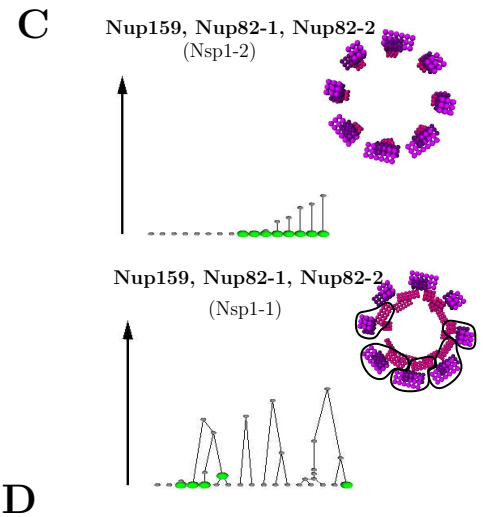


Supplemental Figure 4: Two sub-units of the Y-complex have 16 isolated copies. **Upper-left.** The Hasse diagram of the (Nup120, Nup85, Seh1) sub-complex of the Y-complex exhibits 16 isolated copies appearing at $\lambda = 0$. The red node corresponds to the sub-complex circled on the bottom 3D illustration. **Upper-right.** The Hasse diagram of the (Nup84, Nup145C, Nup133) sub-complex of the Y-complex exhibits 16 isolated copies appearing at $\lambda = 0$. The red node corresponds to the sub-complex circled on the bottom 3D illustration. **Bottom.** 3D model at $\lambda = 0$.

1.4 Further In-silico experiments on Nsp1 related complexes



Protein types	f_{ij} *	k_{high}	k_{drop}	$p_{ij}^{k_{drop}}$	$s(k_{drop})$	$\min \overline{V}_{\lambda_{k_{high}}}$	$\max \overline{V}_{\lambda_{k_{drop}}}$
(Nup49, Nup57)	1	16	16	1.00	1.00	0.47	0.76
(Nsp1-2, Nup49)	1	16	16	1.00	1.00	0.65	0.83
(Nsp1-2, Nup57)	1	16	16	1.00	1.00	0.67	0.86
(Nic96-2, Nsp1-2)	1	6	15	0.86	1.16	0.77	1.82
(Nic96-1, Nsp1-1)	1	7	15	0.82	1.22	0.83	3.66
(Nsp1-1, Nsp1-2)	0.021	1	15	0.63	1.58	0.99	4.63
(Nic96-1, Nsp1-2)	1	6	11	0.93	1.08	0.77	1.51
(Nic96-2, Nup49)	0.442	1	10	0.85	1.18	0.90	2.20
(Nup57, Nup57)	0.005	1	8	0.71	1.41	1.93	2.74
(Nsp1-1, Nsp1-1)	0.021	2	7	0.27	3.72	0.88	8.98
(Nic96-2, Nup57)	0.424	1	3	0.75	1.33	1.22	2.08



Protein types	f_{ij}	k_{high}	k_{drop}	$p_{ij}^{k_{drop}}$	$s(k_{drop})$	$\min \overline{V}_{\lambda k_{high}}$	$\max \overline{V}_{\lambda k_{drop}}$
(Nup159, Nup82-1)	0.951	8	8	1.00	1.00	0.68	0.82
(Nup159, Nup82-2)	0.951	8	8	1.00	1.00	0.65	0.80
(Nup82-1, Nup82-2)	0.284	8	8	1.00	1.00	0.44	0.68
(Nsp1-2, Nup82-2)	1	6	8	0.92	1.09	0.60	0.90
(Nsp1-1, Nup82-2)	1	6	8	0.96	1.04	0.67	0.89
(Nsp1-2, Nup82-1)	1	4	8	0.60	1.66	0.66	2.26
(Nsp1-1, Nup82-1)	1	1	5	0.72	1.39	0.83	3.68
(Nsp1-2, Nup159)	0.187	2	5	0.93	1.08	0.81	1.09
(Nsp1-1, Nup159)	0.187	1	1	0.97	1.08	0.98	0.98

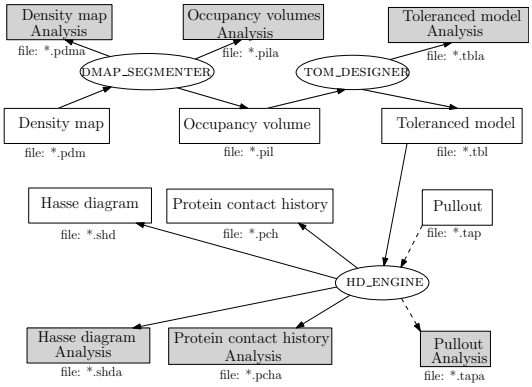
Supplemental Figure 5:

Caption of supplemental Fig. 5

Nsp1 related complexes are of particular interest since distinct fractions of Nsp1 are expected to interact with the T-complex (composed of Nic96, Nsp1, Nup49 and Nup57 [20]) and the Nsp1-Nup82-Nup159 containing complex respectively [21]. Moreover, Nsp1, Nic96, and Nup82 protein types are each represented by two maps in [3]. We used the TOM machinery to infer which sub-population of Nsp1, namely Nsp1-1 or Nsp1-2, is involved in a particular complex.
A. A model of the T-complex and its embedding in the NPC are shown on the left panel, adapted from [20], with only 8 out of the 16 copies presented. The right panel explores the four possible combinations corresponding to the maps (Nic96-1, Nic96-2) and (Nsp1-1, Nsp1-2). For each case, the Hasse diagram shows the isolated copies, from which it is seen that the pair (Nsp1-2, Nic96-2) yields 16 isolated copies, best matching the expectations. For that pair, an instance of the T-complex is presented in the circled region.
B. Contact frequencies f_{ij} from [3] and contact probabilities derived from the TOM for all possible pairs of protein types of the T-complex. Pairs with no contact in the TOM are not represented. For contact frequencies, the * denotes the fact that these frequencies did not discriminate between twin maps, i.e. Nps1-1 and Nsp1-2 on the one hand, and Nic96-1 and Nic96-2 on the other hand. Notice that Nsp1-1 and Nic96-1 do not make contacts with Nup49 or Nup57, strengthening the fact that Nsp1-2 and Nic96-2 likely represent the populations of Nsp1 and Nic96 contributing to the T-complex. The green and orange rows correspond to the six pairs involved in the T-complex with Nsp1-2 and Nic96-2. Note that unlike anticipated [20], only three contacts are observed in the TOM between Nic96-2 and Nup57.
C. The Nsp1-Nup82-Nup159 complex. The 16 copies of Nup82 (8 copies of Nup82-1 and Nup82-2 expected to form dimers [21]), and the 8 copies of Nup159 are asymmetrically positioned on the cytoplasmic side of the NPC. Hasse diagrams and corresponding toleranced models of the 3 types involved in the cytoplasmic Nsp1-containing complex (Nsp1, Nup159, Nup82-1, Nup82-2): the 2 options corresponding to the two Nsp1 maps (Nsp1-1, Nsp1-2) are presented.
D. Contact frequencies f_{ij} from [3] and contact probabilities derived from the TOM for all possible pairs of protein types of the Nsp1-Nup82-Nup159 complex (see also B). Note that while distinct fractions of Nsp1 are expected to interact with the T-complex and the Nsp1-Nup82-Nup159 containing complex respectively [21], Hasse diagrams indicate that, as also observed for the T-complex (A), Nsp1-2 but not Nsp1-1 leads to the formation of the expected 8 isolated copies of this complex.

2 Software: Applications and File Formats

The VORATOM software suite, for Voronoi Analysis of Toleranced Models, is a set of tools meant to design and analyze toleranced models of macro-molecular assemblies. The suite consists of the programs presented on the supplemental Fig. 6, which are all encapsulated within on main executable, namely VORATOM .



Supplemental Figure 6: Overview of the applications: ellipsis represent executable programs; unfilled rectangles represent the main concepts i.e. the objects processed; greyed rectangles represent the analysis associated to instances of the main concepts.

2.1 Overall application

Given a set of maps, the VORATOM building blocks, which are also made available on a stand-alone basis, perform the segmentation of the maps into occupancy volumes (section 2.2), create a toleranced model from these occupancy volumes (section 2.3), and compute the Hasse diagram associated to the toleranced model (section 2.4).

In presenting the executables, we also report the running times obtained on a dual core Intel Extreme CPU X7900 2.80GHz with RAM size of 8Go, running Fedora Core 14. We also note that our programs, written in C++, were compiled with g++ at the optimization level -O3.

- DMAP_SEGMENTER
  The computation of the occupancy volumes of all the 33 maps was done in 28.3 seconds (18 seconds for loading all the maps and 10.3 seconds for the computations). A total of 448 protein instances were reported.
- TOM_DESIGNER
  The computation of the toleranced model of the NPC from the 448 occupancy volumes was done in less than one second. A total of 8064 toleranced balls were reported.
- HD_ENGINE
  As mentioned in the Supplemental, the computation of the protein contact history from the 8064 toleranced balls with $\lambda_{\max} = 1$ failed. We used the partial λ-complex that ran in 72.5 seconds. 2507 protein contacts were reported. The computation of the Hasse diagram from this protein contact history was done in 9.9 seconds.

2.2 Density map segmenter

Consider a map, namely a 3D matrix with one number $\in [0, 1]$ per voxel.

The map segmenter, named DMAP_SEGMENTER in the sequel, selects from the map a prescribed set of connected regions called occupancy volumes.

Input. The main argument is a map. See the .pdm file format on the supplemental Fig. 7.

Output. The main output is a list of occupancy volumes, one per protein instance. A given occupancy volume is represented by the (x, y, z) coordinates of the voxels allocated to this instance, and the density of each voxel is also reported. See the .ovl file format on the supplemental Fig. 8.

Analysis. There are two analysis. The first one, at the protein instance level, compares the occupancy volume against the protein reference volume (the volume estimated from its sequence). The second one, at the map level, compares the number of occupancy volumes created versus the stoichiometry of the protein. (Typically, if the stoichiometry of the protein is larger than the number of connected components of voxels with a non null probability, it may not be possible to create the desired number of instances.)

```
# Global attributes of the map: name of the protein type,
# stoichiometry, number nv of voxels along the x y and z directions
Nup84 16 100
# Cartesian coordinates of the bottom-left corner
x y z

# Densities: nv * nv * nv real numbers, each in the range 0..1
0 0 0 0...
```

Supplemental Figure 7: The .pdm file format to represent a cubic map—the number of voxels is the same along each direction.

```
# Global attribute: total number of occupancy volumes
448

# Then, we find the list of occupancy volumes. Here is one example,
# namely an instance of the protein type Nup192:
# Nup192: protein type;  0: instance index; 240: number of voxels of
# the occupancy volume attributes to this instance
Nup192 0 240
# Then, a list of 240 voxels; for each, the  Cartesian coordinates
# xyz,  and the probability density value
45 41 17 1
...
```

Supplemental Figure 8: The .ovl file format to represent the occupancy volumes of a list of protein instances.

2.3 Toleranced Model Designer

The Toleranced Model Designer, called TOM_DESIGNER in the sequel, computes a toleranced model from a list of occupancy volumes. Note that each toleranced protein consists of a list of toleranced balls.

Input. The main argument is a list of occupancy volumes, each corresponding to one protein instance. (See the .ovl file format.)

Output. The output is the toleranced model. See the .tbl file format on the supplemental Fig. 9.

Analysis. Given a λ value and a list of protein instances, the analysis of a toleranced model consists of computing the volume ratios of Eq. (2). Note that the volumes of these instances are computed amidst the whole NPC. Also, the volume calculation is carried out using affine α-shapes [22], as computing the volume of restrictions in the CW VD is an open problem.

```
# Global attribute: total number of toleranced balls
8064

# Then, a list of toleranced balls. Here is an example toleranced
# ball, represented by the Cartesian coordinates of the center, the
# inner radius, the outer radius, and the index of the protein instance
# this ball belongs to
42.4916 39.2358 16.4461 1.4702 4.19091 0
...
```

Supplemental Figure 9: The .tbl file format to represent the toleranced balls of a list of protein instances.

2.4 Hasse Diagram Engine

The Hasse Diagram engine, called HD_ENGINE in the sequel, computes the Hasse diagram of a toleranced model.

This computation requires two steps, namely the computation of the λ-complex of the toleranced model, and that of the Hasse diagram of the protein complexes. While computing the Hasse diagram, we also store the list of merges between pairs of protein instances, which we call the *protein contact history*. In fact, we successively compute (i) the λ-complex, (ii) the protein contact history, (iii) the Hasse diagram.
The file formats for the Hasse diagram and the protein contact history are presented on the supplemental Figs. 11 and 12 .

Input. While the main argument is the toleranced model, the following options are available:

- One can specify a list of pullouts; if so, the toleranced model manipulated falls into the bicolor setting.
- If a protein contact history is provided, the Hasse diagram is directly derived from it, without computing the λ complex.
- A value $\lambda_{\max}$ can be specified to bound the growth process of the toleranced model. Recall that this value should be set in accordance with the uncertainties observed on the input data, measured by volume ratios.
- Since the computation of the whole λ complex may be time consuming, the partial λ-complex, which consists of the Gabriel simplices of dimension zero and one, may be resorted to. One option is provided to resort to this subset of the λ-complex [23].

Output. If no protein contact history is provided, the one computed from the (partial) λ-complex one is reported. In any case, we report the Hasse diagram involving all protein instances. Furthermore, one finds one Hasse diagram per pullout specified, if any.

Analysis. The following pieces of information are reported:

- The lifetime of the complexes found in the Hasse diagram, and the number of complexes as a function of λ.
- The isolated copies found in the Hasse diagram. (Recall that a pullout is mandatory to define isolated copies.)
- The contact probabilities of protein type pairs. Note that the contact probabilities requires a value for $\lambda_{\max}$. If such a value has not been specified, $\lambda_{\max} = 1$ is used.
- The volume ratio of each complex found in the Hasse diagram is computed, for the λ corresponding to its birth date.

```
# An example pullout.
# A pullout is represented by a triple namely
#(i)   pullout index
#(ii)  the number of protein types in the pullout
#(iii) the list of protein types, the first one being the tagged
#      protein
#
# As an example, here is the pullout of the Y-complex in Sali et al:
54 7 Nup84 Seh1 Nup85 Nup120 Nup145C Sec13 Nup133
...
```

Supplemental Figure 10: The .tap file format to represent a pullout i.e. a list of protein types.

```
# An example protein contact history.
# An element is represented by a triple namely
#(i)   protein instance p1 (type + index)
#(ii)  protein instance p2 (type + index)
#(iii) the weight for which p1 and p2 are connected.
Nup84 55 Nup133 89 0.515
...
```

Supplemental Figure 11: The .pch file format to represent a protein contact history.

```
# Global attribute: pullout, see .tap file format
54 7 Nup84 Seh1 Nup85 Nup120 Nup145C Sec13 Nup133
# total number of vertices and edges in the diagram
495 493

# Vertex description: vertex index and weight of the vertex
66 -0.289302
# Then, description of the protein complex of the vertex:
# number of vertices and edges in
2 1
# list of vertices of the protein complex: pair (type name,
# instance index) of the protein instance in the vertex
Nup133 73
Nup84 114
# list of edges of C: two pairs (type name, instance index) following
by the weight of the edge linking the instances
Nup133 73 Nup84 114 -0.289302

# Edge description: vertex indices of the ancester and the son
66 52
```

Supplemental Figure 12: The .shd file format to represent a Hasse diagram. In the example, the protein complex of the vertex number 52 (not shown) has one protein instance (Nup84, 114).

References: Supplemental

[18] Y. Harpaz, M. Gerstein, and C. Chothia. Volume Changes on Protein Folding. *Structure*, 2:641–649, 1994.

[19] F. Cazals and D. Cohen-Steiner. Reconstructing 3d compact sets. *Computational Geometry Theory and Applications*, 45(1-2):1–13, 2011.

[20] N. Schrader, P. Stelter, D. Flemming, R. Kunze, E. Hurt, and I.R. Vetter. Structural Basis of the Nic96 Subcomplex Organization in the Nuclear Pore Channel. *Molecular Cell*, 29(1):46–55, 2008.

[21] S.M. Nailer, C. Balduf, and E. Hurt. The Nsp1p Carboxy-Terminal Domain Is Organized into Functionnaly Distinct Coiled-Coil Regions Required for Assembly of Nucleoporin Subcomplexes and Nucleocytoplasmic Transport. *Molecular and Cellular Biology*, 21(23):7944–7955, 2001.

[22] F. Cazals, H. Kanhere, and S. Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Transactions on Mathematical Software*, 38(1), 2011.

[23] F. Cazals and T. Dreyfus. Multi-scale Geometric Modeling of Ambiguous Shapes with Toleranced Balls and Compoundly Weighted α-shapes. In B. Levy and O. Sorkine, editors, *Symposium on Geometry Processing*, Lyon, 2010. Also as INRIA Tech report 7306.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399